Coronory Heart Disease Prediction Project code

-Thanojkumar Guntupalli

-Nataliya Volkova

\* Formatting the dataset;

Proc Format;

Value sexft 1 = 'Male'

0 = 'Female'

Other = 'Miscoded';

Value TenYearCHDft

1 = 'CHD'

0 = 'No CHD'

Other = 'Miscoded';

run;

\* Importing the dataset;

Proc Import datafile = "/home/u62112846/Homework/HeartDiseasePrediction.csv"

out = HeartData

DBMS = csv

replace;

getnames = yes;

run;

Data Heart\_Data;

set HeartData;

glucose\_new = input(glucose, BEST32.);

drop glucose;

rename glucose\_new=glucose;

run;

Data Heart\_clean1;

set Heart\_Data;

if cmiss(of \_all\_) gt 0 then

delete;

run;

\*About Dataset variables;

Proc Contents data = Heart\_Data;

Run;

\*Descriptive statistics of variables;

proc means data=heart\_clean1;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*EDA;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*Distribution of TenYearCHD;

Proc Gchart data = heart\_clean1;

Title "Pie chart of Distribution of TenYearCHD";

pie TenYearCHD;

format TenYearCHD TenYearCHDft.;

run;

\* total number of patients by gender;

pattern2 color=brown;

axis1 label=(a=90 "Total number of patients" ) minor=none order= (0 to 2200 by 100) offset=(0,0);

axis2 label=("gender");

Proc gchart data=heart\_clean1;

Title "Total number of patients by gender";

vbar3d sex / discrete width=8 space=7 axis=axis1 maxis = axis2 outside=freq;

format sex sexft.;

run;

\* plot gender against 10 year risk of CHD;

axis1 label=(a=90 "Total number of patients by gender" ) minor=none order= (0 to 2200 by 100) offset=(0,0);

axis2 label=("10 year risk of CHD");

proc gchart data=heart\_clean1;

vbar3d TenYearCHD/discrete group=sex patternid=midpoint width=8 space=6 axis=axis1 maxis=axis2 outside=freq;

format sex sexft.;

format TenYearCHD TenYearCHDft.;

Title "10 year risk of coronary heart disease by gender ";

pattern1 c=green;

pattern2 c=blue;

run;

\* correlation matrix;

proc corr data = Heart\_clean1 plots=matrix;

title "Correlations between age cigsperday totchol sysBP diaBP BMI heartRate glucose ";

var age cigsperday totchol sysBP diaBP BMI heartRate glucose ;

run;

\*smoking vs gender;

Proc Gchart data = heart\_clean1;

hbar currentSmoker /discrete subgroup= sex width=6 space=6 outside=freq;

Title "Proportion of smokers by gender";

format sex sexft.;

pattern1 c=pink;

pattern2 c=blue;

run;

\* freq table prevalentHyp & TenYearCHD;

proc freq data=heart\_clean1;

Title "2X2 freq table of prevalentHyp vs TenYearCHD ";

table prevalentHyp \* TenYearCHD ;

run;

\* box plot of BMI VS TenYearCHD;

proc sgplot data=heart\_clean1;

Title "Box plot of BMI and TenYearCHD";

vbox BMI/group=TenYearCHD;

format TenYearCHD TenYearCHDft.;

run;

\* T-test whether means of BMI of patients with CHD and without are equal or not ;

Proc Ttest data=heart\_clean1;

Class TenYearCHD;

Var BMI;

Title " T-test for BMI";

Run;

\* wilcoxon test;

Proc npar1way wilcoxon data=heart\_clean1;

class TenYearCHD;

var BMI;

run;

\* Distribution of cigsPerDay of people w.r.t their education levels;

proc sgplot data=heart\_clean1;

Title "Box plot of cigsPerDay with respect to education";

vbox cigsPerDay/group=education grouporder=ascending;

run;

\* scatter plot of sysBP vs diaBP;

Symbol value = circle I = none;

Proc gplot data = heart\_clean1;

Title "scatter plot of sysBP vs diaBP";

plot sysBP\*diaBP;

run;

\*Distribution of age of people w.r.t TenYearCHD;

proc univariate data=heart\_clean1 noprint;

Title "Distribution of age";

class TenYearCHD;

histogram age/normal;

inset mean (6.2) var (7.2) skewness (6.2) / Pos = NE;

run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Building the Logistic model;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

/\* Split data into two datasets : 80%- training 20%- validation \*/

Proc Surveyselect data=heart\_clean1 out=split seed= 1234 samprate=.80 outall;

Run;

Data training validation;

Set split;

if selected = 1 then output training;

else output validation;

Run;

ods graphics on;

Proc Logistic Data = training descending plots=all;

format sex sexft.;

class sex( param=ref ref="Female");

class education(param=ref ref="1") currentSmoker( param=ref ref="0");

class prevalentHyp(param=ref ref="0") prevalentStroke(param=ref ref="0");

class diabetes(param=ref ref="0") BPmeds(param=ref ref="0") ;

model TenYearCHD= sex age diaBP cigsPerDay totchol prevalentHyp diabetes BPmeds prevalentStroke education BMI heartRate sysBp glucose currentSmoker / Risklimits lackfit ctable selection = stepwise slstay=0.05 slentry=0.05 stb pprob=0.4;

score data=training out = Logit\_Training fitstat outroc=troc;

score data=validation out = Logit\_Validation fitstat outroc=vroc;

Run;

Quit;

\* scatter plots of Response variable vs significant predictors;

symbol1 value=circle;

proc gplot data=heart\_clean1;

plot TenYearCHD\*age;

Title" scatter plot of TenYearCHD vs age";

run;

quit;

symbol1 value=circle;

proc gplot data=heart\_clean1;

plot TenYearCHD\*cigsPerDay;

Title" scatter plot of TenYearCHD vs cigsPerDay";

run;

quit;

symbol1 value=circle;

proc gplot data=heart\_clean1;

plot TenYearCHD\*sysBp;

Title" scatter plot of TenYearCHD vs sysBp";

run;

quit;

symbol1 value=circle;

proc gplot data=heart\_clean1;

plot TenYearCHD\*glucose;

Title" scatter plot of TenYearCHD vs glucose";

run;

quit;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Building classification Tree model;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

Data heart\_clean2;

set heart\_clean1;

TenYearCHD\_new = put(TenYearCHD, BEST32.);

drop TenYearCHD;

rename TenYearCHD\_new=TenYearCHD;

run;

proc hpsplit data=heart\_clean2 seed=1234 ;

class sex ;

class TenYearCHD;

class education currentSmoker;

class prevalentHyp prevalentStroke;

class diabetes BPmeds;

model TenYearCHD (event='1')= age sex cigsPerDay sysBp glucose diaBP totchol prevalentHyp diabetes BPmeds prevalentStroke education BMI heartRate currentSmoker;

output out= TreeOut;

grow entropy;

prune costcomplexity;

run;